Building a Database of Data Sets for Health Services Research

Sandra J. Frawley, Ph.D.
Center for Medical Informatics and Department of Epidemiology and Public Health
Yale University School of Medicine, New Haven, CT

The Database of Data Sets (DB/DS) for Health Services Research will be an online searchable directory of data sets which are available, often with restrictions and confidentiality safeguards, for use by health care researchers. The DB/DS project is aimed at a wide audience, and intends to include a very broad range of health care data sets, ranging from state hospital discharge data bases, to national registries and health survey data sets, to institutional clinical databases. The intended users are the same community of researchers, policy-makers, administrators and practitioners who are served by the National Library of Medicine's current bibliographic databases. This paper describes a pilot phase of the DB/DS project in which the issues involved in creating such a database were explored with an initial set of 20 representative data sets.

INTRODUCTION

A Database of Data Sets (DB/DS) for Health Services Research is under development by the National Library of Medicine through its recently created National Information Center on Health Services Research and Health Care Technology. Growing out of work on the Unified Medical Language System Information Sources Map, the DB/DS will be an online searchable directory of data sets which are available, often with restrictions and confidentiality safeguards, for use by health services researchers.

With the establishment of the Agency for Health Care Policy and Research (AHCPR) in 1989, Congress set out as national priorities 1) the conduct of medical effectiveness research, and 2) the creation of clinical practice guidelines based on the results of past, present, and future research. The importance of bringing such information to clinical practice is an important theme of the health care reform proposals being considered in 1994.

Recognizing that time, resources, and sometimes ethical considerations will frequently not permit the use of prospective clinical studies to explore the efficacy of clinical treatment and patterns of care, various organizations, including the Institute of Medicine, have urged that researchers be given better information about and access to previously collected data (1). It is likely that research funding will be increasingly available for secondary analysis of data rather than for primary data collection and analysis. As a result, there is a major need to store information about such data sets in an organized fashion so that it can be searched easily by diverse researchers.

Health services researchers, broadly defined, span the spectrum from social scientists who study the organization and financing of medical care to clinicians who evaluate the relative effectiveness of alternative medical treatments. Among the social scientists, there is historically a greater tendency to utilize national or state data sets that have been collected by organizations other than the researchers themselves. With research questions focusing more on community patterns of disease and the outcomes of individual diseases and procedures, epidemiologists and clinical researchers have been less interested in the use of previously collected data. Rather, the emphasis has been on prospective, primary data collection efforts and using randomized clinical trials to evaluate treatments. Recently, however, there has been extensive discussion and debate about the feasibility and desirability of using existing administrative and clinical data sets for researching medical effectiveness issues (2-5).

The DB/DS project is aimed at a wide audience, and intends to include a very broad range of health care data sets, ranging from state hospital discharge data bases, to national registries and health survey data sets, to institutional clinical databases. Data sets sponsored or created by government agencies, private foundations, private corporations, and health care institutions will be included. The intended users are the same community of researchers, policy-makers, administrators and practitioners who are served by the National Library of Medicine's bibliographic databases such as MEDLINE, Health Planning and Administration, and Health Services/Technology Assessment Research (HSTAR).

A central issue in building the DB/DS is determining how best to code (catalog) the various data sets to serve the needs of health services researchers.

0195-4210/94/$5.00

Researchers looking for secondary sources of data often start by asking where they can find information about a particular population they want to study, e.g., elderly Hispanics with congestive heart failure. As a result, the coding of a data set has to provide an accurate description of its content: the unit of analysis (the entities about which data has been collected) and the variables (the data items collected). The researcher also needs to understand the methodology of data collection and the quality of the data produced, both of which will affect the suitability of the data set for the researcher's purposes. In many instances, the researcher may want to combine data from several sources, so issues of linkability among data sets become important. Technical issues such as the formatting of data files and the electronic media in which the data set is available may determine whether the researcher will be able to analyze the data at his/her institution. Finally, he or she needs to understand the data access procedures, e.g., any restrictions that apply, from whom the data set can be acquired, and the likely cost.

This paper describes a pilot phase of the DB/DS project in which these issues were explored with an initial set of 20 representative data sets.

RELATED WORK

The need for flexible access to data sets is widely acknowledged. Two examples of systems which have approached this problem are described below.

The Centers for Disease Control and Prevention (CDC) recently made 24 public health databases available for on-line searching and data analysis through CDC WONDER (6). This system contains various data sets from national health and hospital discharge surveys as well as textual and bibliographic databases such as the Morbidity and Mortality Weekly Report. CDC WONDER users can search with keywords and be directed to the relevant databases. Because the data are stored online, users have immediate access to the information and can download selected data into personal computers.

On a much larger scale, though with less user search and analysis capability, are the data set archives developed by the Inter-university Consortium for Social and Political Research (ICPSR) at the University of Michigan (7). Started over thirty years ago, the archives now contain about 600 gigabytes in more than 30,000 machine readable data files, including a substantial number of health-related data sets. Users may search an online directory of ICPSR data sets using a SPIRES interface or a gopher server. A small number of frequently requested files are stored on a Sun workstation and made available to users by network file transfer protocol. Most data sets must be purchased in magnetic tape, CD-ROM, or diskette format.

Several differences between these programs and the NLM's new database are important to recognize. The NLM does not plan to archive coded data sets, nor put them online. Instead, detailed information about the contents, format, purpose, source, restrictions on use, and availability will be maintained online in a searchable database. Many data sets that will be coded are proprietary. Others exist as institutional information systems, from which data would have to be extracted for research purposes.

SELECTING DATA SETS FOR THE DB/DS PILOT PROJECT

During the pilot phase of the DB/DS project described in this paper, the goal was to develop an efficient yet robust coding scheme that would capture the information researchers will want to know about a data set, using a pilot group of 20 data sets. We wanted the 20 data sets chosen for initial coding to be representative of the great variety that will eventually be contained in the database. As a result, the main emphasis was on selecting data sets with different contents in terms of the units of analysis and types of information collected. Not surprisingly, the pilot group also showed wide variety in size, sponsorship, methodology, and accessibility.

Although the DB/DS project is not designed to be an archive of data sets, it became clear that accurate coding requires that a hard copy, or machine-readable copy, of the data set documentation be available to the coder. As a result, it is necessary to assemble this material as a reference tool during the coding process even though the actual data files themselves will not be acquired. Data sets that are created by federal government agencies and sold through HCFA, the NTIS or CDC tend to have extensive documentation available for users, including data dictionaries, data instruments, codebooks and tape layout descriptions. This material frequently includes the instructions originally given to the data collection team. Other data sets were established for institutional use and without the expectation that external researchers might utilize the data. These data sets often do not have documentation that can be

made readily available to external users, even when there exists internal documentation appropriate for the needs of institutional users.

The listing below shows the group of 20 data sets coded for the pilot study.

Clinical Records
1. Duke DataBank for Cardiovascular Disease (Duke University Medical Center, Durham, NC)
2. HELP System Data Base (LDS Hospital, Salt Lake City, UT)

Discharge Summaries
3. Healthcare Cost and Utilization Project, 1988-1994 (AHCPR)
4. National Hospital Discharge Survey, 1988 (NCHS)
5. NY State Hospital Discharge Data Files, 1991 (NY State Department of Health)

Claims Records
6. MEDSTAT Market Scan Data Base (MEDSTAT Systems)
7. Quality Care MEDPAR File, 1988 (HCFA)
8. Standard Analytical Inpatient Public Use File, 1992 (HCFA)
9. Standard Analytical Physician/Supplier Public Use File, 1991 (HCFA)

Epidemiological Surveys
10. Tecumseh Community Health Study, 1959-1969 (Victor Hawthorne, et al.)

Health/Behavioral/Social Surveys
11. National Medical Expenditure Survey, 1987 (AHCPR)
12. National Health Interview Survey, 1988 (NCHS)
13. National Long Term Care Survey, 1982-1984 (DHHS)
14. National Survey of Access to Medical Care, 1982 (Ronald M. Andersen and Lu Ann Aday)

Disease Registries
15. ARAMIS - Arthritis, Rheumatism, and Aging Medical Information System (American Rheumatism Association)
16. Cancer Surveillance and Epidemiology in the United States and Puerto Rico, 1973-1977 (National Cancer Institute)

Birth Registries
17. Natality Detailed Data File, 1988 (CDC)

Data About Practitioners
18. AMA Physician Master File, 1992 (American Medical Association)

Data About Programs and Facilities
19. National Evaluation of Rural Primary Health Care programs, 1979-1982 (Cecil G. Sheps and Edward H. Wagner)
20. National Nursing Home Survey, 1985 (NCHS)

CODING THE DATA SETS

The main purpose of the pilot project was to explore the issues that arise in coding a representative group of data sets, and to develop an initial coding strategy which will be refined over time as the project evolves. In the coding scheme developed during the pilot project, 42 fields are available to describe the data sets, although not all fields are used for each entry. Some are free text, while others require that selections be made from pre-defined choices. The basic data provided for each data set include its name and any alternative names, the source (creator), a general description of the data set and the purpose for its creation, the person who prepared the entry, and the date of the entry.

Content of the Data Set: Roughly half of the fields deal with the content of the data set. After this content is summarized in a free text general description field, a series of fixed choice fields provide more precise information: whether the data set uses a standard vocabulary or coding scheme, such as DRG or CPT; whether it contains a standard or minimum set of variables, e.g., UB-82; the unit(s) of analysis; the setting(s) of data collection; the timeframe; geographic locations and level of detail; age groups included; ethnic groups included; and a listing either of individual variables (if fewer than 25) or of categories of variables.

Methodology: Methodology fields provide information on the universe, the sample size, sampling procedures, and mechanisms of data collection. A free text field allows for brief comments on methodology issues.

Published References: Two fields are intended to assist researchers in using the published literature to learn about a particular data set. The REFerences field gives citations to publications describing the data set, while STUdies contains citations to publications which have analyzed the data set. In addition, an

eventual goal is for the DB/DS to contain a unique identifier for each data set that can be published in research reports and included in NLM database citations.

Accessibility: Accessibility fields describe the restrictions which may be placed on data use and access and also list address and cost information about the providers (vendors) of the data set.

Linking Data Sets: Many data sets are produced as part of annual or periodic surveys and other data collection efforts, e.g. the National Hospital Discharge Survey, the National Health Interview Survey, and the Natality Detailed Data File. Common data elements appearing over time facilitate longitudinal investigations. Other data sets can be linked together because the same personal identifier, e.g., patient number or physician license number, is used, as in various Medicare files. Probabilistic matching algorithms are being developed to permit data file linking even when common identifier numbers are not used (8). When links to other data sets are known, they can be described in the Related Entries field.

Technical Information: Information about technical documentation, the editing of data, the quality of data, the size, format, and structure of electronic data are provided in technical information fields. As described previously, the availability of technical information in the written documentation is quite variable.

CODING ISSUES THAT AROSE

In this section we discuss certain problematical issues that arose in developing a coding strategy for the pilot group of data sets.

Quality of the Data

After questions of content, perhaps the most critical issue for researchers is the quality of the data. This has proved to be the most difficult to capture within the coding scheme because there is usually not very much discussion of data quality in written documentation. In addition, quality may be highly variable from item to item within a data set. Researchers will be able to draw some inferences about quality from coded information about sampling and data collection methods.

An excellent source of information is likely to be the principal investigator or research director in the organization which created the data set. Over time however, this information becomes more difficult to retrieve. Personal memories fade and data producing organizations rarely invest many resources in documenting the decisions associated with data collection and production (9). Similarly, other researchers who have used the data set could have useful insights into data problems and limitations. It would be an interesting challenge to create a vehicle within this proposed database for researchers to communicate their discoveries about a data set.

Coding Tried and Discarded, and Why

An early version of the DB/DS coding included a field to indicate the likelihood that a data set could be used for each of 17 types of health services research. The different types of health services research, ranging from treatment effectiveness and outcomes studies to cost and utilization assessments, were based on discussions in recent publications of the Institute of Medicine, the NLM, and the AHCPR. We soon recognized, however, that it would be difficult for a data set coder to make a definitive judgment about the suitability of a data set for a particular purpose, and that researchers themselves have the responsibility of making such determinations. Furthermore, we were aware that in the dynamic field of health services research, 17 brief phrases were not likely to be a satisfactory statement of the breadth and depth of research concerns.

We also did not attempt to rate data sets according to the AHCPR's medical effectiveness criteria for database content (10). Two of these criteria are that treatment information and outcome information be included. There is a vast difference in the amount and usefulness of treatment or outcome data which appear in a discharge abstract, for example, and that which might appear in a computer-based patient record or observational study record. Rather than make judgments about the adequacy of treatment and outcome information, it appeared more reasonable to use the Variables or Variable Categories field to describe what treatment- and outcome-related data items actually appear in the records.

DISCUSSION

In performing this pilot analysis, we have attempted to look at a range of different data sets that illustrate a variety of issues that need to be confronted both in coding and in using the DB/DS. A number of interesting issues arose, as discussed below.

Great Variation in Number of Data Items
There were extreme differences in the number of variables included in different data sets. A few, e.g., Cancer Surveillance, contained only a handful of variables which may be handled by explicitly listing those variables. Other data sets, e.g., National Long-Term Care Survey, contained thousands of variables which therefore had to be summarized. Since even exact variable names can obscure the meaning of a data item, e.g., "Claim Non-Covered Day Count," coders will not always be accurate in describing the content of data items when summarizing large numbers of variables. As a result, facilitating user access to documentation, such as data dictionaries, codebooks and instruments, would be very helpful. We are exploring the possibility of making documentation available on the Internet via FTP so that researchers could immediately learn data definitions and see the actual wording of questions.

Differences in Scope and Accessibility
Many data sets, e.g., National Hospital Discharge Survey, exist as a public use file, readily available for purchase. Other data sets, e.g., the AMA Physician Masterfile, are very large files, from which researchers typically request a subset dealing with a particular specialty or region. Still other data sets, e.g., HELP, are large dynamic clinical information systems, from which a customized subset of data would be extracted with appropriate safeguards for patient confidentiality and with institutional approval.

Through the data set coding, users of the DB/DS should be informed about special procedures they must follow in order to gain access to institutional data. For example, external researchers have entered into collaborative efforts with Duke researchers to utilize the DataBank for Cardiovascular Disease.

Single vs. Multiple File Structures
While many data sets exist as a single file, others have multiple data files, each with a different unit of analysis and format, but all linked together because the information was gathered as part of a single research effort. To accommodate complex data sets such as the National Health Interview Survey, the coding scheme has the flexibility to allow differentiation of the content and technical descriptions of the various parts.

SUMMARY

Our experiences in the pilot phase of database development have suggested that coders will find great variation in the information they will have to work with in coding data sets. The coding scheme is evolving to be highly, even redundantly, descriptive, but not to require evaluative judgments by coders. We cannot anticipate all the inquiries and methods that health services researchers will bring to this database, and we are content to let users make their own judgments after following the leads we can offer.

Acknowledgements: This work was supported in part by NIH contract N01 LM13537 from the National Library of Medicine. The author would like to acknowledge the participation of Betsy L. Humphreys and Marjorie A. Cahn of the National Information Center on Health Services Research and Health Care Technology at the National Library of Medicine, for whom this work was performed.

References

[1]. Harris-Wehling J, Morris LC, Eds.: Improving Information Services for Health Services Researchers: A Report to the National Library of Medicine. (Washington, D.C.: Institute of Medicine), 1991.
[2]. Kane RL, Lurie N: Appropriate effectiveness: A tale of carts and horses. QRB 18(10):322-6, 1992.
[3]. Grady ML, Schwartz HA, Eds: Medical Effectiveness Research Data Methods. USDHHS, Public Health Service, AHCPR Pub. No. 92-0056, July, 1992.
[4]. Statistics in Medicine 10(4), 1991, entire issue.
[5]. International Journal of Technology Assessment in Health Care 6(2), 1990, entire issue.
[6]. Freide A, Reid JA, Ory HW: CDC WONDER: A comprehensive on-line public health information system of the Centers for Disease Control and Prevention. American Journal of Public Health 83:1289-94, 1993.
[7]. Inter-university Consortium for Political and Social Research: Guide to Resources and Services, 1993-1994, (Ann Arbor, MI: ICPSR), 1993.
[8]. Roos LL, Wajda A: Record linkage strategies. Part I: Estimating information and evaluating approaches. Methods of Information in Medicine. 30(2):117-23, 1991.
[9]. David MH: Systems for Metadata: Documenting Scientific Databases. SSRI Workshop Series, 9221, (Madison, WI: University of Wisconsin), 1992.
[10]. AHCPR: Feasibility of Linking Research-Related Data Bases to Federal and Non-Federal Medical Administrative Data Bases. USDHHS, Public Health Service, AHCPR Pub. No. 1991-0003, April, 1991.